Probability (Practice Problems)

Problem 1:

Suppose you are given the following data:

	company	model	rating	type
0	ford	mustang	A	coupe
1	chevy	camaro	B	coupe
2	ford	fiesta	C	sedan
3	ford	focus	A	sedan
4	ford	taurus	B	sedan
5	toyota	camry	B	sedan

Find the marginal probabilities of the different ratings.
Also find out the conditional probabilities of different type of cars given different ratings of the car.

Problem 2:

Context

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

Content

Attribute Information:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values. One file has been "processed", that one containing the Cleveland database. All four unprocessed files also exist in this directory.

To see Test Costs (donated by Peter Turney), please see the folder "Costs"

Acknowledgements

Creators:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Donor:
David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

Inspiration

Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

See if you can find any other trends in heart data to predict certain cardiovascular events or find any clear indications of heart health.

i. Find the probability of a patient having fasting sugar level > 120 mg/l given that the patient is a male.

ii. Find the probability of a patient having fasting sugar level > 120 mg/l given that the patient is a female.

iii. Find the probability of a patient having serum level 200 – 239 (moderately elevated) given that the patient is male?

iv. Find the probability of a patient having serum level 200 – 239 (moderately elevated) given that the patient is female?

v. P (hypertension | older than 60) =